3) Generation of both soft and hard copy displays of data contained in the Context Base and the MDB.

4) User selectable analysis of any data available to the DMS.

The implementation of METASYS is not yet complete. The PMON and the basic MVI processes are functional, but the DMS development has not yet been completed. This development is proceeding rapidly, however, which can be attributed to the advantages of program development on the PDP 10.

1.9 Summary

As the above hardware and software improvements are being made we will continue evaluation of the GC/HRMS system in parallel with its actual application to real problems. GC/HRMS is a relatively new and difficult technique for routine application. In order to use it effectively, we will have to exert some effort toward determining and optimizing the performance of the many elements of the system, the GC, the MS, and the computer hardware and software.

2 PART 2: DEVELOPING PERFORMANCE AND THEORY FORMATION PROGRAMS TO ASSIST IN BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS

2.1 Introduction

The Heuristic DENDRAL computer programs assist with structure elucidation problems by helping interpret mass spectra and helping generate structures that are consistent with data obtained from a variety of other spectroscopic and physical/chemical sources. The Meta-DENDRAL programs assist with rule formation problems in cases where the rules of mass spectrometry are not known.

Both the interpretation and rule formation programs are written as interactive tools to be controlled by professionals to combine the professional's judgment with the computer's combinatorial power.

2.2 CONGEN

The CONGEN [48,53] program represents a significant extension of a program which has developed over the last several years, the cyclic structure generator [40,41]. The purpose of CONGEN is to assist the chemist in determining the chemical structure of an unknown compound by 1) allowing him to specify certain types of structural information about the compound which he has determined from any source (e.g., spectroscopy, chemical degradation, method of isolation, etc.) and 2) generating an exhaustive and non-redundant list of structures that are consistent with the information. The generation is a stepwise process, and the program allows interaction at every stage; based upon partial results the chemist may be reminded of additional information which he can specify, thus limiting further the number of final structures.

CONGEN fits with the other DENDRAL programs as a "backstop" solution to structure elucidation problems. If the mass spectrum of an unknown compound is available, then CLEANUP and MOLION could be used, but if the general class of the compound is not known, PLANNER has no starting point from which to work. In such cases, structural information can be extracted manually from the spectrum and given to CONGEN for analysis. Because CONGEN makes no assumptions about the source of this information, other spectroscopic or chemical techniques may be used to supply supplemental data.

At the heart of CONGEN are two algorithms whose accuracy has been mathematically proven and whose computer implementation has been well tested. The structure generation algorithm [31,37,40,41] is designed to determine all topologically unique ways of assembling a given set of atoms, each with an associated valence, into molecular structures. The atoms may be chemical atoms with standard chemical valences, or they may be names representing molecular fragments ("superatoms") of any desired complexity, where the valence corresponds to the total number of bonding sites available within the superatom. Because the structure generation algorithm can produce only structures in which the superatoms appear as single atoms (we refer to these as intermediate structures), a second procedure, the imbedding algorithm [48,53] is needed to expand the superatoms to their full chemical identities.

These two routines give the chemist the ability to construct

structures from a given set of molecular "building blocks" which may be atoms or larger fragments. By itself, this capacity is of limited utility because the number of final structures can be overwhelming in many cases. Usually, the chemist has additional information (if only some general rules about chemical stability, which the program has no concept of) that can be used to limit the number of structural possibilities. For example, he

may know that because of a compound's stability, it cannot contain a peroxide linkage (O-O) and thus the programs need not consider such structures when there are two or more oxygens in the "building block" list.

In the past year CONGEN has reached the level of a practical production program which can aid chemists, both locally and at remote network sites, in solving the structures of drug-related compounds and natural products. The development of this program during the year has been strongly guided by the difficulties and new requirements which have appeared as it was applied to a wide variety of cases, and its efficiency and usefulness have increased dramatically. We report here the details of the modifications and additions we have made to CONGEN, and the effects they have had on its utility. Also, because of the rich repertoire of structure modification and testing functions available within CONGEN, we have found it to be an invaluable "laboratory" for the testing of new ideas, and we briefly describe two pilot projects which form the basis for future research. Discussion of applications of CONGEN to problems of biochemical interest is included in Part 3.

Program modifications

DEPTH-FIRST GENERATION. This modification has been both the most difficult and the most useful. The structure-generation algorithm which was originally part of CONGEN processed the "tree" of subgoals and subgoals-of-subgoals in a breadth first fashion. Although this was the most logically coherent and understandable encoding of the algorithm, it meant that a user would have to wait until the very end of a generation problem before he could see any of the results. This was particularly frustrating when a problem was submitted to CONGEN which was too big and/or time-consuming, because the user could never get any results at all. To alleviate this difficulty, we undertook a complete reorganization of the structure-generation algorithm so that it would proceed depth-first, giving results continuously as the computation progressed.

It is difficult to communicate the complexity of such a reprogramming without a major digression, but the flavor of the necessary changes is captured in the following example. At several points in the algorithm, there are what might be called "branching functions" whose purpose it is to solve some intermediate problem which has several alternate solutions. It is easiest to define such a function so that it computes the whole list of possibilities and returns the list to the caller. It is then the caller's responsibility to determine what is to be done with each possibility, and the branching function itself can be viewed as a separate module. This is a breadth first approach, and the difficulty is that the caller can make no progress until the branching function has constructed and returned all possibilities. The depth-first approach is to have the branching function itself be responsible for further

processing each time it creates a new result. To retain the modularity of the branching function, some mechanism is needed to allow the caller to "tell" it what this further processing consists of, and such a mechanism was instituted throughout the structure-generation algorithm.

We made use of the depth-first generation by instituting an interrupt mechanism in CONGEN whereby a user can examine the developing list of structures as they are created. This is a tremendous advantage both psychologically, because it gives the user a feeling that the program is "doing something", and operationally because it provides rapid feedback. A chemist can now often see quickly that a given case will create many more structures than expected, and the intermediate output can suggest forgotten constraints or superatoms. The following is an example of a terminal session in which the interrupt mechanism is used. The character control-S gives a "snapshot" of progress on the problem while control-I allows for the drawing of partial results. Both of these features are illustrated in the sample CONGEN session shown in Appendix A.

NEW CAPABILITIES FOR THE USER. There have been several additions to CONGEN which are visible to the user and which generally increase the flexibility and power of the program. These include

1) Making CONGEN aware of aromaticity, a chemical property of molecules which results from certain combinations of double bonds in rings. Aromaticity has a profound effect upon both the chemical reactivity and symmetry properties of molecules, and CONGEN can now be directed to detect aromaticity in its output structures, to compensate for the difference between the actual symmetry of an aromatic system and the symmetry which appears in the graph representing it, and to distinguish aromatic from non-aromatic atoms when it tests GOODLIST and BADLIST entries.

2) Giving the user the ability to type "?" to any prompt in the program, which results in a summary of the possible inputs. In some cases this summary is a list of possible commands, while in others it is a short explanatory message. A new interactive teletype-input routine was developed which makes it easy to include such help messages in the program, and which mimics the handy command-recognition and command-completion features of the TENEX operation system.

3) Including new specifications in the EDITSTRUC language for describing substructural features. The user can now declare a bond in a substructure to be an "anybond", which means that the atoms at the termini are connected but that the multiplicity of the connection is unspecified. This is especially handy when defining substructures containing aromatic portions because bond multiplicity is an indistinct concept in aromatic systems. Another new structural element which can be specified is a "linknode", a node which stands for a variable-length chain of

atoms of the given type rather than a single atom. The minimum and maximum lengths of such a chain can be specified as well. The linknode feature is useful for defining constraints on ring fusions and other constraints such as Bredt's rule which depend on path length. Other extensions have been made internal to CONGEN which will shortly be reflected in the user-level language of EDITSTRUC. These include numerical inequalities involving node properties (e.g., "the number of H's on atom 3 is greater than the number of H's on atom 5") or linknode lengths (e.g., "the sum of the lengths of linknodes 2 and 6 is greater than 5"), and greater control over the number of fittings found for a GOODLIST constraint (e.g., the ability to distinguish between "the number of N's in six-membered rings" and "the number of six-membered rings containing N").

4)Allowing greater flexibility in the selection of terminal type. This choice controls the output of structural drawings so they are best suited to the user's terminal. Several different types of character-oriented and graphics-display terminals are now supported.

5) Making CONGEN accessible from the GUEST login account at SUMEX. This involved preventing a GUEST user from reaching certain critical points in CONGEN which would allow greater system access than is normally authorized for guests. We can now offer trial access to CONGEN via the guest mechanism without worrying about SUMEX misuse.

6) Creating a BATCH command for CONGEN. This allows the user to submit time-consuming, compute-bound calculations to the batch-processing facility of SUMEX. The computation is then run automatically at off-hours when it will not overload the system resources. The user can now run CONGEN in its interactive mode to input all of his data and then submit the large tasks to BATCH for overnite processing.

7) Including a pruning function MSPRUNE which is used to test a list of candidate structures for consistency with a set of observed peaks from a mass spectrum. The candidates are typically generated by CONGEN using structural data from other sources. The user specifies the observed MS peaks (as elemental compositions or nominal masses or a combination of both) along with a set of constraints on the allowed cleavage processes. MSPRUNE retains only those candidates which can account for the observations via one of these allowed processes. The constraints speak of the number of bonds broken and the number of steps in a process, the proximity of pairs of cleaved bonds (i.e., whether or not two adjacent bonds can break in a given process), the multiplicity or aromaticity of each cleaved bond and the possible neutral transfers. MSPRUNE is the first CONGEN function which can aid directly in the interpretation of "raw" spectral data.

8) Internal CONGEN Developments. The basic algorithms used for structure generation in CONGEN are firmly rooted in

mathematical graph theory. During the past year, there has been significant refinement of several of these graph theoretical algorithms. The new algorithms have been coded in SAIL, an extended ALGOL type language; and a sophisticated executive has been developed to coordinate the various SAIL routines as well as to direct the communication and control between the SAIL component and LISP component of CONGEN.

The power and utility of CONGEN rests, to a great extent, on the fact that it can generate structures under user supplied constraints. The most powerful of the routines used in constrained structure generation is the fragment imbedder [37,48]. It is this routine which permits CONGEN to efficiently generate only those structures containing given polyatomic fragments (i.e., superatoms). The fragment imbedding program was completely rewritten so that it operates now in a "depth first" rather than "breadth first" style. This was done so that the user can request CONGEN to produce examples only of candidate structures in those cases where the total number of candidate structures is very large. This change also increases the efficiency of the fragment embedding process and has the advantage that if a CONGEN run must be interrupted, the user is left with at least some candidate structures rather than just intermediate results.

During the grant period, a very general substructure matching algorithm was developed and coded in SAIL. This algorithm accepts as input a structure and a "pattern" and returns the number of times the pattern distinctly occurs in the structure. Here a pattern is a partially specified substructure in which atom names, bond widths and hydrogen attachments all may assume a range of values. This routine is used by CONGEN for post checking of structures and classifying lists of structures.

An improved technique to determine the topological symmetry group of a structure was also developed and coded in SAIL. This routine is used in several parts of CONGEN, e.g., fragment imbedding. This new routine is, statistically, at least an order of magnitude faster than the old group finding routine.

The language LISP, although quite powerful, does not produce very efficient machine code. It was for this reason that several of the routines used by CONGEN were coded in SAIL. However, because of the widely variant data types, LISP and SAIL are not compatible languages. Hence, all of the SAIL programs reside in their own TENEX fork, and they communicate with the LISP fork via a shared memory page. The new CONGEN SAIL code executive program handles all interfork communication for the SAIL routines, and it allows one to make additions or modifications to the SAIL portion of CONGEN with relative ease. This ease of change is also aided by the fact that all the SAIL programs are written in highly modularized form.

Preliminary testing of the new CONGEN SAIL fork indicates

these modifications and additions will yield a significant increase in the overall efficiency of CONGEN, and hence will enable one to consider a broader range of chemical problems.

INTERNAL CONGEN IMPROVEMENTS - LISP. Because of the divers assortment of chemical problems to which CONGEN has been applied, we have been able to exercise all parts of the program in a variety of contexts. As a result, we have been able to uncover a number of hidden inefficiencies in the LISP section of CONGEN, and although correcting these has not had a direct impact on the command structure of the program, we estimate that a decrease of over 50 in CPU time has been achieved for typical CONGEN cases. In some cases this decrease is as high as 90.

These improvements have been numerous, but one stands out as most significant. Several changes were made to the graph-matching routine which is responsible for testing the presence or absence of structural features in molecules or molecular fragments. The new routine uses list space (a key resource in the LISP programming system) much more parsimoniously, and it incorporates a new and very efficient representation of substructures which makes optimum use of the linked-list data representation in LISP. Also included were a number of heuristics which, although they do not alter the output of the graph matcher, do dramatically decrease the amount of time spent on typical tests. The highly efficient SAIL graph matcher, described above, will soon supplement the LISP version, thought the latter will still be needed in some cases because of its greater flexibility.

Other inefficiencies were detected and fixed in the portion of CONGEN which builds tree-like molecules and molecular fragments, where it was discovered that a built-in assumption (that the most common monovalent atom would be hydrogen) was adversely effecting the running times of some CONGEN cases, and in the portion responsible for computing the symmetry groups of graphs.

PILOT PROJECTS. CONGEN provides an excellent environment for the testing of new ideas because it contains an extensive "library" of functions for the creation, manipulation and testing of topological representatives of molecular structure. Below we describe two pilot projects which were explored within this environment and which provide the basis for proposed future research topics.

We developed within CONGEN a program called XMECH [60] whose purpose it was to study the possible mechanisms of cyclizations and skeletal rearrangements of monoterpanes, terpanes and sesquiterpanes. The study of these compound classes is an important sub-field of natural-products chemistry, and simple carbonium-ion mechanisms, such as cyclizations to double bonds and 1,2-alkyl and/or 1,2-hydride shifts, are frequently invoked to rationalize interrelationships between various

skeletal types. Using XMECH we were able to explore various combinations of these basic mechanisms and to develop exhaustive lists of skeletal types, known and unknown, which should be accessible from known biogenetic precursors via this approach. Our results indicate that although such mechanistic rationalizations are widely used, the method is quite non-selective: If a sufficient number of mechanistic steps is included to account for even a modest fraction of known skeletons, a vastly larger number of skeletal types are obtained which have never been seen in nature. It seems clear that there are much subtler mechanistic considerations which account for the specificity of biogenetic pathways, and our work points out the danger of rationalizing that specificity with an overly simple model. XMECH has laid the groundwork for a much more general program, REACT, in which a user will be able to define chemical reactions and apply them to problems of mechanistic chemistry and structure elucidation.

A second pilot project is the program MDGGEN which embodies a new, general approach to the interpretation of a mass spectrum in terms of structural possibilities for an unknown. The method used in MDGGEN compliments the MSPRUNE function described above (section 7 of NEW CAPABILITIES FOR THE USER) because it uses MS data at the beginning of a problem rather than as a final filter on candidate structures. Whereas MSPRUNE is logically part of the TEST phase in the traditional DENDRAL scheme of PLAN-GENERATE-TEST, MDGGEN logically belongs in the PLAN phase. Conceptually, MDGGEN is related to the PLANNER program, except that MDGGEN analyzes MS data without relying upon class-specific fragmentation rules as does PLANNER. Using a very simple and general fragmentation theory, MDGGEN processes selected peaks from a mass spectrum and constructs possible ways of segmenting the overall composition of the molecule to account for those peaks. These segmented descriptions are graphs similar to topological chemical structures except that one node may stand not just for a single chemical atom, but a collection of atoms (a composition) representing a connected piece of the molecule. We call these mass-distribution graphs, or MDG's. The structure-generation facilities of CONGEN allow us to assemble the atoms within each node-composition in all unique ways, and to imbed these assemblies in all unique ways into the overall MDG structures. In this way, we arrive at chemical structures which account for the MS data according to the simple theory. MDGGEN is still in its infancy, with the practical limitations of computer time and storage requirements restricting it to small molecules (up to perhaps ten non-hydrogen atoms) and relatively few observed peaks (up to roughly seven or eight ion compositions). This early development, which could take place rapidly because of the existing facilities within CONGEN, has helped us to focus our attention on the critical advances which will be needed in creating a more flexible and generally useful program.

2.3 PLANNER

The DENDRAL PLANNER program [28,33] is designed to analyze the mass spectrum of a compound or of a mixture of related compounds. Because there is no **ab initio** way of relating a mass spectrum of a complex organic molecule to the structure of that molecule, PLANNER requires fragmentation rules for the class of compounds to which the unknown belongs. This is its major limitation.

Applications and limitations of PLANNER have been discussed extensively. [28,33] The program is very powerful in instances where mass spectrometry rules are strong (i.e., general, with few exceptions). In instances where rules are weak or nonexistent, additional work on known structures and spectra may yield useful rules to make PLANNER applicable (see INTSUM and RULEGEN, below). One unique feature of

PLANNER is its ability to analyze the spectra of mixtures in a systematic and thorough way. Thus, it can be applied to spectra obtained as mixtures when GC/MS data are unavailable or impossible to obtain.

The power of the PLANNER has been substantially increased by including the MOLION program (discussed below) as a subroutine for computing the list of plausible molecular ions. Since this subprogram does not depend on knowledge of the compound class, the PLANNER no longer needs to have class-specific rules for determining the mass and empirical formula of the unknown molecule.

The major use of the Planner in the past year has been as a means of testing new class-specific mass spectrometry rules proposed by the Meta-DENDRAL program described below. One measure of quality of a set of proposed rules is their ability to discriminate among isomers in the same class. For example, the monoketoandrostane rules can be partly evaluated by their ability to assign the keto group to the correct substituent position, based on the mass spectrum of the compound. Since there are eleven possible positions, we are asking the rules to discriminate the correct structure from the other ten monoketoandrostanes.

2.4 Meta-dendral Rule Formation Programs

When the mass spectrometry rules for a given class of compounds are not known, the INTSUM, RULEGEN and RULEMOD programs can help a chemist formulate those rules. Essentially, these programs categorize the plausible fragmentations for a class of compounds by looking at the mass spectra of several molecules in the class. All molecules are assumed to belong to one class whose skeletal structure must be specified. Also, the mass

spectra and the structures of all the molecules must be given to the program.

INTSUM collects evidence for all possible fragmentations (within user-specified constraints) and summarizes the results. For example, a user may be interested in all fragmentations involving one or two bonds, but not three; aromatic rings may be known to be unfragmented ; and the user may be interested only in fragmentations resulting in an ion containing a heteroatom. Under these constraints, the program correlates all peaks in the mass spectra with all possible fragmentations. The summary of results shows the number of molecules in whose spectra there is evidence for each particular fragmentation, along with the total (and average) ion current associated with the fragmentation.

The INTSUM program [34] is in routine, production use to assist in interpretation of the mass spectra of new classes of molecules (see Part 3 for details).

The RULEGEN program attempts to explain the regularities found by INTSUM in terms of the underlying structural features around the bonds in question that seem to "drive" the fragmentations. For example,

INTSUM will notice significant fragmentation of the two different bonds alpha to the carbonyl group in aliphatic ketones. It is left to RULEGEN to discover that these are both instances of the same fundamental alpha-cleavage process that can be predicted any time a bond is alpha to a carbonyl group.

The RULEMOD program modifies and condenses the set of rules produced by INTSUM and RULEGEN together. It looks at the negative evidence associated with each candidate rule in order to select the best ones, then merges rules that seem to explain the same breaks (if possible). The program was substantially improved in several ways, as described in the next section.

2.4.1 Improvements Made to the Meta-DENDRAL Programs

2.4.1.1 INTSUM Improvements

Transfers of arbitrary neutral species can now be specified as part of the mass spectrometry processes, instead of transfers of hydrogen atoms alone. This capability increases the utility of the program in at least two ways: first, it allows a chemist to control the program better -- to produce the kinds of results that are more chemically meaningful -- and second, it allows the program to explore more complex processes within its space and time limitations. For example, carbon monoxide and water were listed as plausible neutral molecules to transfer in or out of fragments for the triketoandrostanes. Thus, the processes are listed with and without these transfers, just as chemists prefer,

instead of showing loss of CO as a set of two breaks around the keto group, or loss of H_2O as loss of oxygen (breaking the C=O bond) accompanied by loss of two hydrogens. What is more, the program can now produce these results without violating its chemical heuristics of (a) not breaking adjacent bonds, and (b) not breaking double bonds. This economy also pays off in increasing the complexity of the processes that can be considered. Because loss of CO, for example, is a result of a transfer instead of the result of breaking two bonds, the number of bonds broken in accompanying processes can be increased by two.

Another INTSUM improvement was to increase the options for initial data filtering. Thresholding is too simple for many problems, so we now provide an option to cluster peaks and select the n largest peaks from each cluster.

The format of the input data is also now less strict than before. We have written programs to read spectra in Aldermaston format. And we have merged CONGEN's Editstruc package into the INTSUM setup routines to allow a chemist to associate structures with spectra interactivity. This greatly decreases the chances of error in setting up the input data.

Several modifications were also made to the program to increase its efficiency, **e.g.**, processing all intensities as integers (between 0 and 1000).

2.4.1.2 RULEGEN Improvements

The evaluation of prospective rules in RULEGEN guides the entire rule generation procedure. To tune this procedure, we modified the evaluation function in several ways and compared the resulting sets of rules. We were looking for an objective way of telling the program to keep rules general, but "not too general". The current evaluation function is substantially improved as a result.

Because the RULEGEN program searches such a large space of partial and complete rules, it requires large amounts of computer time (sometimes more than 60 cpu minutes). Thus, we have investigated several improvements for efficiency alone. In addition, we have made the program easier to set up and run in batch mode to reduce the chemist's personal time investment. And we have made the program easily restarted from any intermediate point -- to protect the chemist from machine failures.

2.4.1.3 RULEMOD Improvements

At the time of the last annual report RULEMOD was a new prgram still in its experimental stages. Since then we have added new subprograms and integrated the program with other programs to make it a useful and necessary part of Meta-DENDRAL.

Two new subprograms greatly improve RULEMOD's performance. (1) A program to add specifications to rules was completed. It looks for plausible ways of making a rule more specific in order to decrease the number of counterexamples to the rule. (2) A complementary program to make rules more general was also completed. The program tries to find ways to reduce the number of descriptors on nodes of subgraphs in order to increase the breadth of applicability of rules. Its major constraint is that it cannot make any change that would increase the number of counterexamples. Both of these subprograms make the final rules much closer to rules that chemists approve of.

The subprogram that merges rules was also improved. The program tries to merge pairs of rules into a more general form for economy and clarity of rules. Its major constraint is that no explanations are lost, **i.e.**, all the data points explained by the initial pair of rules will still be explained after merging. Formerly we insisted that the more general form must cover all the same data points as the initial rules, but this was found to be too narrow a constraint. By giving the program a more global view of the entire set of rules, we can let the more general, merged form explain fewer data points than its component rules as long as other rules explain the remainder.

2.4.2 Search for New Applications of the Rule Formation Programs

In this year the Meta-DENDRAL programs have matured enough to let us consider extending them beyond mass spectrometry. The domain that we chose was 13C NMR spectroscopy, for a variety of reasons.

13C NMR has been characterized as the spectroscopic technique of the 1970's [68]. Our laboratories have been involved in experimental work on 13C NMR spectra of amines, keto and hydroxy steroids [62-64]. In addition, we have carried out a preliminary investigation of a Heuristic DENDRAL approach to interpretation of 13C spectra of amines [39].

There are several parallels between rule formation in mass spectrometry and 13C NMR spectrometry. In both techniques the precise reasons for molecular fragmentation (in the former) or NMR absorption (in the latter) are poorly understood. In the absence of a detailed theory capable of accurate prediction of spectra, we seek empirical rules which can relate observed data to measurable structural parameters. Some of the structural parameters presumed relevant, **e.g.**, atom type, bond multiplicities, are shared in both techniques. Some of the current Meta-DENDRAL structural manipulation functions can be used for either technique. An important difference is that the planning phase of Meta-DENDRAL (**i.e.**, INTSUM) necessary in applications in mass spectrometry is not required for 13C NMR because we will deal initially with spectra whose absorption

peaks (or "shifts" relative to any internal standard) are assigned to specific atoms in the known structures. Typically scientists have sought an explanation for the 13C NMR shift of an atom in terms of the structural environment of the atom. Searching such structural environments is a problem which is amenable to solution by existing and proposed parts of the Meta-DENDRAL program.

As in applications to mass spectrometry [58] we will propose a set of factors which might affect 13C NMR absorptions. With a description of these factors we will use the Meta-DENDRAL program to produce a set of rules which will reproduce and predict resonance shifts of individual 13C atoms.

The current Meta-DENDRAL program represents a basic framework for studying 13C NMR rule formation. We believe that the program will require little revision to accommodate the differences in data and rules. We have already considered some of the problems of changing the form of rules. The subgraphs in the situation parts of rules need to be generated "outward" from a specific 13C atom instead of outward from a bond broken in the mass spectrometer. The action parts of rules need to take account of an explicit absorption range whereas for mass spectrometry the rules predict much more precise data points (mass positions). We have made a preliminary test of the program's extensibility in the context of alkanes.

For the alkane study we used only a topological model of molecular structure, not a geometric model. The rules that were formed from a test set predicted shifts for 13C atoms in other alkanes (outside the test set) with accuracy within 1.5 ppm. The major modifications needed in the program to produce these preliminary results were the following:

(a) change RULEGEN to generate rules by expanding the subgraph environments outward from a central atom rather than from a central atom rather than from a central bond;

(b) change the form of rules to associate a range of shifts with each subgraph rather than a precise fragment mass;

(c) redefine RULEGEN's evaluation function for partial rules to take account of the desire to predict narrow ranges of shifts.

Other domains were considered, including finding rules to associate pharmacological activity with molecular structure and finding rules for other organic chemical analysis techniques. Of all that we considered, 13C NMR appears to offer the most in terms of both feasibility and utility.

2.5 Results

2.5.1 Keto-androstanes

We have shown that the Meta-DENDRAL program is capable of rationalizing the mass spectral fragmentations of sets of molecules in terms of substructural features of the molecules. On known test cases, aliphatic amines and estrogenic steroids, the Meta-DENDRAL program rediscovered the well-characterized fragmentation processes reported in the literature. On the three classes of ketoandrostanes for which no general class rules have been reported, the mono-, di-, and triketoandrostanes, the program found general rules describing the mass spectrometric behavior of those classes. The general rules shown in Tables II, IV, and VI explain many of the significant ions for compounds in these classes while predicting few spurious ions. The program has discovered consistent fragmentation behavior in sets of molecules which have not appeared by manual examination to behave homogeneously in the mass spectrometer.

Programs with knowledge of the scientific domain can provide "smart" assistance to working scientists, as shown by the reasoned suggestions this program makes about extensions to mass spectrometry theory. We are aware that the program is not discovering a new framework for mass spectrometry theory; to the contrary, it comes close to capturing in a computer program all we could discern by observing human problem-solving behavior. It is intended to relieve chemists of the need to exercise their personal heuristics over and over again, and thus we believe it can aid chemists in suggesting more novel extensions to existing theory. It can be argued that the two-dimensional connectivity model of molecules used in this study is not the right model for mass spectrometry; that there are deeper rationalizations of a fragmentation process than subgraph environments. However, this model is commonly used by working chemists and once fragmentations based on this model are defined, chemists can readily provide the remaining "mechanistic" rationalizations or see that further experimental work with labeled compounds is necessary. (Other limitations of the method have been discussed at the end of the methods section.)

Recent statistical pattern recognition work addresses some of the points on rule formation and spectrum prediction raised in this paper. We have avoided blind statistical methods for three important reasons. 1) We wish to explore thousands of possible subgraphs with associated features, as we search for those which are in some way important. Current pattern recognition procedures are restricted to much smaller numbers of manually (or computer-assisted) selected features, adding additional bias to the procedure. 2) We want to know how certain rules were obtained by the program and why certain other rules were rejected or not detected. We can trace the reasoning steps of the Meta-DENDRAL program and determine chemically meaningful answers to

such questions in a way that is not possible with purely statistical programs. 3) We wish to constrain the rule formation activity in ways that are natural to a working chemist. For example, we may want the program to avoid fragmentations involving aromatic rings or two bonds to the same atom, or, as mentioned above, we may want to look at fragmentations accompanied by loss of CO or other neutral fragments.

Rules can be formulated to explain data in terms that are known to be meaningful to chemists; most importantly, the rule formation constraints are under the control of the chemist. Also we feel that this approach provides a high level of generality in describing fragmentation processes. Although the rules are developed in the context of a particular set of compounds, they are not tied to that set but can be applied in other contexts, or compared to rules developed from other sets of compounds in a search for common features of the rules. For these reasons, we believe that the Meta-DENDRAL program offers a powerful and useful complement to pattern recognition programs for finding relationships between structures and spectral data.

We are cautiously optimistic about the general applicability of this rule formatton method, although we have demonstrated its utility for only a small number of compound classes and only in the context of mass spectrometry.

2.6 Heuristic Programming Project Workshop

In the first week of January, 1976, about fifty representatives of local SUMEX-AIM projects convened at Stanford for four days to explore common interests. Six projects at various degrees of development were discussed during the conference. They included the DENDRAL and META-DENDRAL projects, the MYCIN project, the Automated-Mathematician project, the Xray-Crystallography project, and the MOLGEN project. Because of the interdisciplinary nature of each of these projects, the first day of the conference was reserved for tutorials and broad overviews. The domain-specific background information for each of the projects was presented and discussed so that more technical discussions could be given on the following days. In addition the scope and organization of each of the projects was presented focusing on the tasks that were being automated, how people perform these tasks, and why the automation was useful or interesting.

In the following days of the workshop, common themes in the management and design of large systems were explored. These included the modular representations of knowledge, gathering of large quantities of expert knowledge, and program interaction with experts in dealing with the knowledge base. Several of the projects were faced with the difficulties of representing diverse kinds of information and with utilizing information from diverse

sources in proceding towards a computational goal. Parallel developments within several of the projects were explored, for example, in the representation of molecular structures and in the development of experimental plans in the MOLGEN and DENDRAL projects. The use of heuristic search in large, complex spaces was a basic theme to most of the projects. The use of modularized knowledge typically in the form of rules was explored for several of the projects with a view towards automatic acquisition, theory formation, and program explanation systems.

For each of the projects, one session was devoted to plans for future development. One of the interesting questions for these sessions was the effect of emerging technology on feasibility of new aspects of the projects. The potential uses of distributed computing and parallel processing in the various projects were explored, particularly in the context of the DENDRAL project.

Most of the participants felt that the conference gave them a better understanding of related projects. And because many members of the SUMEX-AIM staff actively participated, the workshop also provided all projects with information about system developments and plans. The discussions and sharing of ideas encouraged by this conference has continued through a series of weekly lunches open to this whole community.

3 PART 3: APPLICATIONS TO BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS

3.1 Introduction

In our grant proposal we discussed the application of the instrumentation and computer programs described above to the study of molecular structure problems in a variety of biomedical applications areas. This is our primary research area, and we discussed specific classes of problems and compounds for investigation. We also made it quite clear that our facilities would be made available to wider community of collaborators/users as our resources permitted. Both categories of application, i.e., within our own group, and with an outside group, are described in some detail below. Our last annual report described several steps taken to encourage a broad community of researchers to use our facilities. For example, we sent a questionnaire to members of the American Society for Mass Spectrometry, Committee

III on Computer Applications, and a follow-up letter to persons indicating a desire to know more about access to our programs. The same note has been sent to several other persons whom we know from personal contacts might be interested. Because of the nature of their investigations, many of these people receive NIH support. Several of our publications (e.g., [45-49,53-61]) mention the availability of our programs. In addition, through individual contacts and formal presentations at conferences we have been encouraging outside use of the programs.

The availability of SUMEX as a mechanism for resource sharing has made it possible for us to extend access to our programs to a number of people. Without SUMEX, this access would be impossible, and most of our programs (those which are not easily exportable) could be used only by ourselves.

3.2 Applications by Professor Djerassi's Research Group

Our existing grants, outlined below, mesh well with our instrumentation and program development under the present award. Under NIH Grant GM06840 we have been studying natural products from marine sources with major emphasis on terpenoids and sterols. For this work we have been dependent on the use of our 711 instrument for high resolution mass spectrometry which we require for the identification of all new compounds, many of which are present in only very small quantities. We were particularly anxious to have access to GC coupled with a high resolution mass spectrometer because we hope to be able to screen large numbers of marine animals for their sterol content using this technique. We are currently engaged in intensive efforts in analysis of mixtures of marine sterols involving our computer-based procedures. The program for the development of the computer operated and assisted system of marine sterol structure analysis has been planned to proceed in three stages:

1) Analysis of all literature published concerning marine sterols so that a complete listing of known sterol structures and organisms studied could be compiled.

2) Collection, evaluation, digitization and computer file construction for the mass spectra of all known marine sterols, followed by the institution of a computer operated file search sequence for direct analysis of marine sterol GC-MS data.

3) The application of the INTSUM, RULEGEN, and RULEMOD programs to the computer file of marine sterol spectra so that a series of fragmentation rules can be extracted for use in the generation of possible structures from mass spectral data for new marine sterols, that is, sterols whose mass spectra cannot be matched with any spectra contained in the computer search file.

We are presently completing the second stage and beginning the third. The following discussion will be a summary of the work that has been completed, and the work that is in progress or planned.

The literature concerning marine sterols is extremely extensive. Over a thousand reports concerning marine sterols can be found scattered throughout a multitude of journals dating back to the initial report by Henze in 1908. In spite of the occurrence of a number of good review works in the literature, we have found the compilation of all reported marine sterol structures and organisms studied to have been an imposing task, which we have now completed successfully. The search has also pointed up a number of entire phyla of marine invertebrates for which no sterol analysis have been reported, and has therefore pointed out perhaps the best candidates to which the developing automated analytical procedures should be applied. The search has also generated an extensive and very refined list of descriptions which are now used in a computer generated update of our bibliography every two weeks for this very active field. This laboratory has been involved in sterol work for some years and so our own samples and mass spectral files have made a significant contribution to the compilation of the complete mass spectral file of marine sterols.

Table I represents a listing of marine sterol spectra as well as a listing of purely synthetic sterol mass spectra (for use in evaluation of the INTSUM results) which have been contributed by this laboratory. These spectra are now part of completely functional computer files. We have requested and received samples of other marine sterols from researchers around the world who have reported their isolation. A large number of these sterols have now had mass spectra taken and the enlargement of our computer mass spectral file is proceeding rapidly.

The series of programs for processing raw GC-MS data and searching mass spectral files have recently been instituted on the chemistry PDP 11/45 computer. The series of programs which have potential application to processing our data are CLEANUP (a program for subtracting GC column bleed or background and noise from raw GC-MS data, and resolving spectra of overlapping elutants), MOLION (a program for generation of molecular ion candidates from mass spectral secondary losses), and SEARCH (a program for searching and comparing experimental mass spectra to the file of known marine sterol mass spectra). Several data management programs exist for displaying the results of the file search and other operations. Development of a program to utilize GC retention indices is progressing. The first experimental file search for an actual sample run will be possible within the next few weeks, but we have already used the SEARCH program to process and evaluate several duplicate marine sterol mass spectra from our files as listed in table I. Table II represents the results of this kind of experiment. Three separate (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL (trivial name "FUCOSTEROL") mass spectra

were compared to 25 marine sterol mass spectra in the computer files via the SEARCH program. The program was able to select each of the mass spectra from the main file with the inclusion of one thirty carbon sterol (24Z)-24-PROPYLIDENECHOLEST-5-2N-3BETA-OL which possesses a structure similar to FUCOSTEROL, the twenty-nine carbon sterol. This kind of study has shown that in principle the SEARCH program functions for marine sterol correlations, but requires some fine tuning to reduce this kind of error. The search strategy modifications should be complete within the next several weeks.

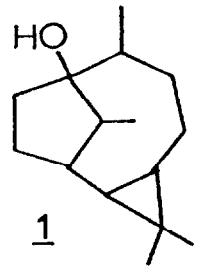
One other aspect of this work should be mentioned. We have found that for very complex marine sterol mixtures a single GC-MS run is sufficient to identify the major sterol components and a few minor components. Further separation procedures are required to analyze the remaining minor components. We have found many of the minor components to be of significant biosynthetic and ecological interest. We have spent a considerable effort perfecting rapid separations or enrichments of these minor sterol components so that GC-MS analysis can be run on them. We now have a procedure utilizing silica gel, alumina , silver nitrate impregnated alumina and silica gel, and high pressure reversed phase liquid chromatography which produces separations and/or enrichments so that GC-MS data can be obtained for every sterol of even a 30 component mixture. Perfecting these separations have required over six months. We have used the sterol extracts of two Gorgonians or soft corals, Pseudoplexaura Porosa and Plexaura Homomolla. Within these extracts we have discovered several new classes of marine sterols, including several twenty-two carbon sterols of unusual stereochemistry, a twenty-one carbon sterol, several new 5-BETA stanols, and a series of extremely interesting 19-nor-delta-5-sterols (publications in preparation). We feel certain that with the institution of the computer assisted procedures described herein, the time required for this kind of study (half a year) can be cut down to weeks.

Application of INTSUM to the marine sterol spectral files has just begun. One aspect of the INTSUM work which should be mentioned here is that in addition to the free 3-beta-hydroxy marine sterol files, a number of marine sterol derivatives (acetates, O-methyl ethers, trimethylsilyl ethers, and other derivatives) were compiled from the mass spectral library in this laboratory. INTSUM will be applied to these marine sterol derivative files in order to extract fragmentation rules. Comparison of the results for the free and derivatized sterols will point up the cases where some of the derivatives (which have superior GC properties) can be used with a minimum of loss of mass spectral information. We are confident that the file search system will be functioning before July. We already have marine extracts arriving from our collaborators in Brazil, and have offered the use of the system, once it is functioning, to researchers in Japan and Britain. We feel that the system will be of great benefit to the large number of researchers in the marine sterol field.

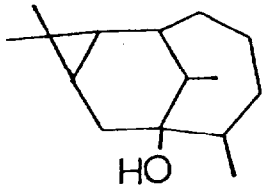
Another major area of interest in our chemical laboratories is the structural analysis of marine terpenoids using CONGEN in conjunction with a variety of spectroscopic data collected on these compounds. For the past year we have been involved in the application of CONGEN in the area of structural elucidation specifically related to marine natural products other than steroids. CONGEN's advantages in these studies lie chiefly in its ability to provide interactively the chemist with assurance that no plausible solutions have been overlooked, as well as an insightful measure of the progress of the problem, thereby suggesting clues to guide the course of the investigation.

(+)-Palustrol.

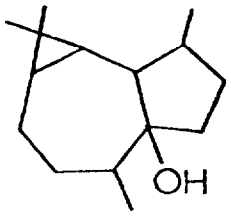
The utility of CONGEN has been demonstrated recently [57] in the identification of (+)-palustrol, a tricyclic sesquiterpene alcohol from the marine **Xeniid Cespitularia virdis.** Inferences derived from 1H and 13C nmr spectra suggested molecular fragments whose assembly by CONGEN resulted in an initial set of 272 candidate structures. Examination of the set suggested appropriate nmr decoupling experiments resulting in the imposition of additional constraints which reduced the initial set of candidates to 88. Dehydration of the tertiary alcohol and spectral examination of the resulting olefins provided additional structural constraints which reduced the set further to 22. Recognition of an additional constraint after examining these possibilities eliminated two of the 22. Of the remaining 20 structures, only four (1 - 4) obey the isoprene rule, and of these four, 1 and 2 may be deleted because there dehydration would yield unsaturated analogs which violate Bredt's rule.



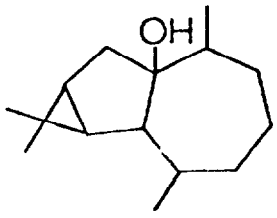
1



2



3



4

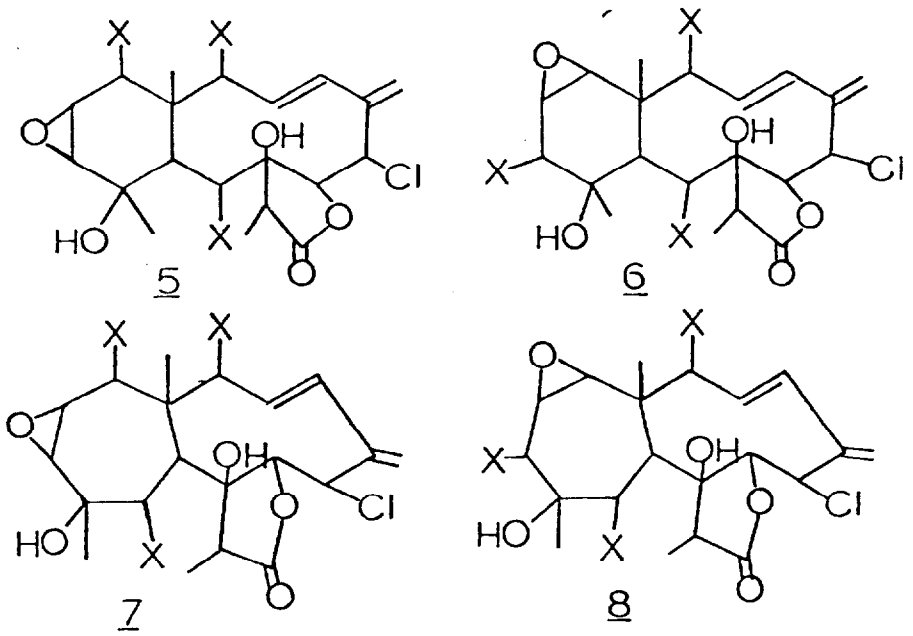
Examination of the literature revealed that structure 3 had been assigned to (-)-palustrol (L. Doleijs, V. Herout, and F. Sorm, CCCC, 26, 811 (1961)). The published infrared spectrum for

(-)-palustrol was identical in all respects to that of the unknown alcohol, thus establishing its structure. Our structure 3, however, displays the opposite rotation of polarized light.

Briareine D.

A recent study by Tursch and Bartholome (C. Bartholome, PhD. Thesis, University of Brussels, 1974) resulted in two alternative proposed structures for Briareine D, one of four chlorinated diterpene lactones isolated from the gorgonian **Briareum asbestinum.**

Rigorous examination of the structural inferences which led to the proposed structures yielded molecular fragments and constraints which were supplied to CONGEN for construction of structural candidates. The results confirmed the proposed structures, 5 and 6, and, more importantly, suggested two additional candidates (7,8) which had not been considered previously and could not be excluded on the basis of existing data.



(X = RCOO)

Work is currently in progress on the CONGEN-assisted structure elucidation of the aglycone portion of Lemnalialoside, a diterpene glycoside from **Lemnalia digitata**, and a tricyclic sesquiterpene hydrocarbon from **Sinularia mayi.**

Further applications are summarized under headings of subsequent sections which refer to specific programs. Much of the effort in application of our programs to the mass spectral data implicitly assumes that the data are available. In fact, without the current and future instrumentation effort discussed in Part 1, these program applications would not be feasible.

3.2.1 CLEANUP

The spectral cleanup program, written for ourselves and our collaborators in the Dept. of Genetics, Stanford Hospital (see Local/Stanford Community, below) is now in routine use. A manuscript describing the method is now in press [61]. Several improvements have been made in the program to increase its capabilities for dealing with complex multiplets of overlapping GC peaks and to improve its efficiency. The resulting version of the program has been exported to several other laboratories which have expresses interest in our methods (see end of Part 3).

3.2.2 INTSUM

As a means of extending the rules of fragmentation in mass spectrometry, several classes of compounds are under study as we attempt to determine characteristic modes of fragmentation. The following is a brief description of each such class and the current status of our research:

1. Pregnanes: Pregnanes related to the progesterone skeleton have been analyzed in some detail in collaboration with Dr. S. Hammerum, (University of Copenhagen, Denmark). Two manuscripts describing this work have recently appeared [65,66].

2. Androstanes: Keto-substituted analogs of the skeleton of the important steroidal hydrocarbon, androstane, were being studied in collaboration with Dr. Roy Gritter (an IBM scientist who spent his sabbatical leave in our laboratory learning more about mass spectrometry). This study is important to our understanding of the mass spectral behavior of complex, polycyclic systems. It is providing a model for the use of Meta-DENDRAL programs. We have completed this study and a manuscript describing our method and results is now in press [58] in the Journal of the American Chemical Society.

3. Macrolide Antibiotics : We have finished the first stages of our analysis of the fragmentation of several members of these macrocyclic systems. We have solicited and obtained a mall number of additional compounds to supplement our own limited number of samples. We are currently correlating the INTSUM results from closely related structures to identify systematic modes of fragmentation. We are designing experiments of deuterium labelling and metastable defocusing to help distinguish among alternative explanations by INTSUM for several prominent ions in the spectra of these compounds. Further efforts on this problem are hindered by lack of available standards.

4. Insect Juvenile Hormones: In collaboration with Dr. Loren Dunham, Zoecon Corp., we are investigating regularities in

the fragmentation behavior of the juvenile hormones. Previous work on the mass spectra of these compounds was carried out only at low resolving powers. We have obtained the high resolution mass spectral data for these compounds and have completed the INTSUM analysis of the data. Our findings have been described in a manuscript which will appear shortly [67] in Organic Mass Spectrometry. Our results will prove valuable for structural analysis and detection of these compounds and congeners.

5) Marine Sterols : The previous section summarizes our continuing efforts in marine sterol analysis, including the importance of INTSUM in these studies.

3.2.3 RULEGEN AND RULEMOD

As described above, RULEGEN and RULEMOD can be used to assist in discovery of mass spectrometry fragmentation rules which depend on substructural features of molecules. Thus, it can be used for classes of compounds where the fragmentation does not depend on the basic skeleton, but on local features expressed by common substructures. Our studies [58] on the performance of the program (see Meta-DENDRAL section) have involved analysis of spectra of previously well-characterized classes of compounds. We have analyzed spectra of aliphatic amines and estrogenic steroids in terms of fragmentation dependence on substructural features of these molecules. Excellent agreement with literature descriptions of fragmentation were obtained. We then proceeded with a study of the previously uncorrelated mono-, di- and triketoandrostanes. Our results [58] provide new insights into regularities of molecular fragmentation among members of the same group. The results also indicate little or no additivity of effects of keto substitution; spectra of diketoandrostanes are not superpositions of the respective monoketoandrostanes.

3.2.4 CONGEN

We are currently engaged in efforts to explore the utility of CONGEN to a variety of structure elucidation problems. The current areas of application are summarized below, together with progress to date.

1) Ion Structures: CONGEN has been used to construct possible ion structures under a variety of constraints in support of studies on the structures of ions in the mass spectrometer. These studies are crucial to a deeper understanding of molecular fragmentation. The programs results are used to ensure that no plausible alternatives have been overlooked during efforts to characterize the structures. We have recently published a detailed description of the use of CONGEN which illustrates the systematic approach available with the program [55].

2) Terpenoid Systems: We are using CONGEN to explore questions of the scope of terpenoid isomerism. We would like to determine some criteria which might allow us to say something about why only certain structural types are found in nature, to the exclusion of many possibilities which are very similar in structure. A manuscript describing our first results is now in press in Tetrahedron [60] and describes some aspects of the structural isomerism of mono- and sesquiterpenoid skeletons.

3) Scope of Structural Isomerism: We are investigating the philosophical and pedagogical aspects of the scope of structural isomerism. This investigation is important to our program design and strategy as we identify the ways persons consider and reject whole categories of structural possibilities. A manuscript describing this work has appeared in the Journal of Chemical Information and Computer Science [54].

4) Constraint Implementation: A detailed description of the kinds of constraints available to guide CONGEN in its exploration of structural possibilities has been presented [56]. This description also presents how constraints and efficient implementation of chemical "common sense" were derived from considerations of manual approaches to structural problems.

5) Marine Natural Products: The previous section described use of CONGEN in solving unknown structures in this area of application of our techniques.

3.3 Utilization of the Mass Spectrometry Resource

3.3.1 Applications of High Resolution Mass Spectrometry

A) Prof. Djerassi's Group

We have run about 75 samples to obtain high resolution mass spectra in support of DENDRAL research problems. These have included marine sterols (acquisition of reference spectra and verification of structures of new synthetic materials), macrolide antibiotics, ketoandrostanes and substituted pregnanes for Meta-DENDRAL studies of fragmentation processes.

B) Stanford Chemistry Department.

We have run a number of spectra for other researchers in the Department of Chemistry. Samples have included a number of diterpanes, alkaloids and unknown compounds from both chemical and enzymatic cyclization procedures.

C) Other Stanford Community

We have run spectra for a number of our collaborators in the Medical School. These have included samples from the Departments of Genetics, Psychiatry and Anaesthesia, representing structural analyses of metabolic products, drug purity and possible reaction products of an anesthetic, respectively.

D) U.S. and Foreign Collaborators.

Spectra have been obtained for Dr. Dunham, Zoecon Corp., of Juvenile hormones for INTSUM studies [67]; Dr. Gritter, now back at IBM, steroids for Meta-DENDRAL studies [58]; Dr. Fitch, Yale University, alkaloid metabolites; Dr. Tomer, Univ. of Brooklyn, spectra for fragmentation studies; Dr. Jaeger, Univ. of Wyoming, structure identification of crown ether components; Dr. Spangler, Univ. of Idaho, structure identification of sulfides for studies of remote sulfur-sulfur interaction in the mass spectrometer. High resolution spectra have been provided to Dr. Nakano, Venezuela, alkaloids, Drs. Mors and Gilbert, Brazil, steroids and alkaloids, Dr. Sultanbawa, Ceylon, triterpenes and alkaloids, and Dr. Orazi, Argentina, terpenoids.

3.3.2 Applications of GC/High Resolution Mass Spectrometry.

During the past year we have analyzed the following samples by GC/ HRMS (these samples represent real applications and do not include the many samples of standard compounds which were analyzed during this time during development of the GC/HRMS system):

A) Prof. Djerassi's group - We have analyzed about 40 mixtures of marine natural products, primarily sterols, by GC/HRMS. Some samples were standard compounds necessary as reference materials but available only as mixtures. Some samples were mixtures of unknown compounds. Spectra were obtained primarily on underivatized sterols, occasionally from acetate derivatives.

B) Other Stanford collaborators - We have run GC/HRMS analyses of several mixtures of diterpenes and precursors, and enzymatic and chemical cyclization products of squalene epoxide analogs for Prof. van Tamelen, Dept. of Chemistry. We have analyzed ten urine fractions in conjunction with on-going work with Prof. Lederberg's group in the Dept. of Genetics. These have been primarily organic and amino acid fractions, derivatized as appropriate, and urinary polyamines analyzed as the trifluoroacetate derivatives.

3.3.3 Other Mass Spectral Studies

We have obtained a number of conventional mass spectra (low resolution) in cases where high resolution data were not required